Network Based Analysis of Cancer Data

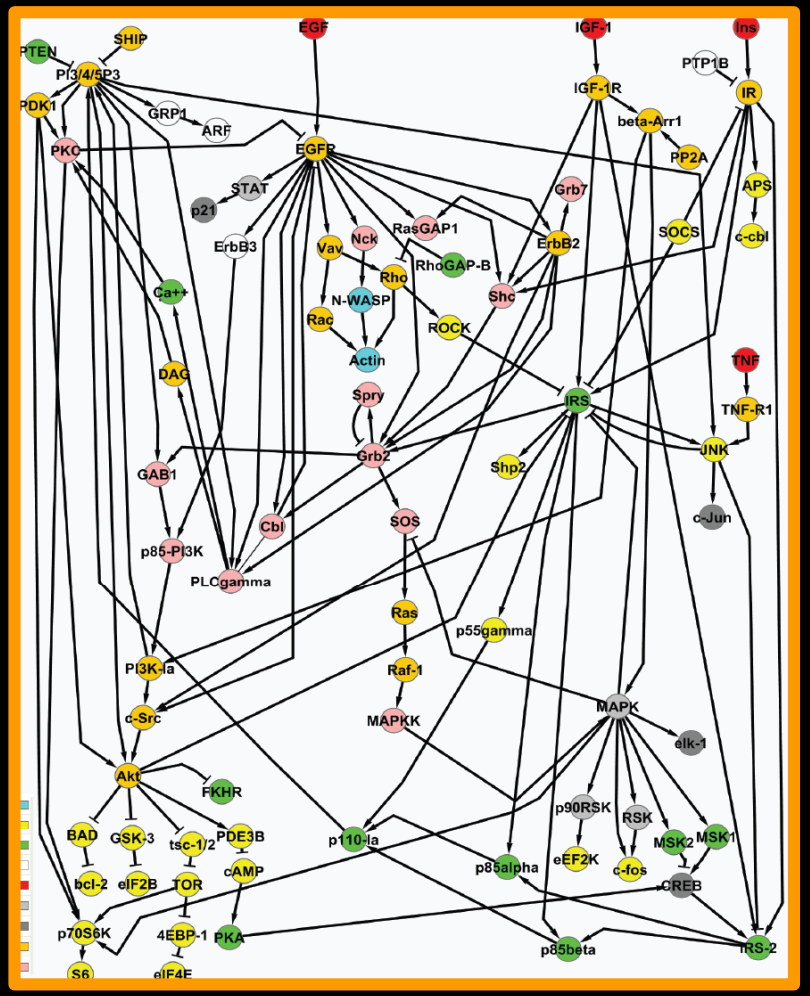
Teresa Przytycka

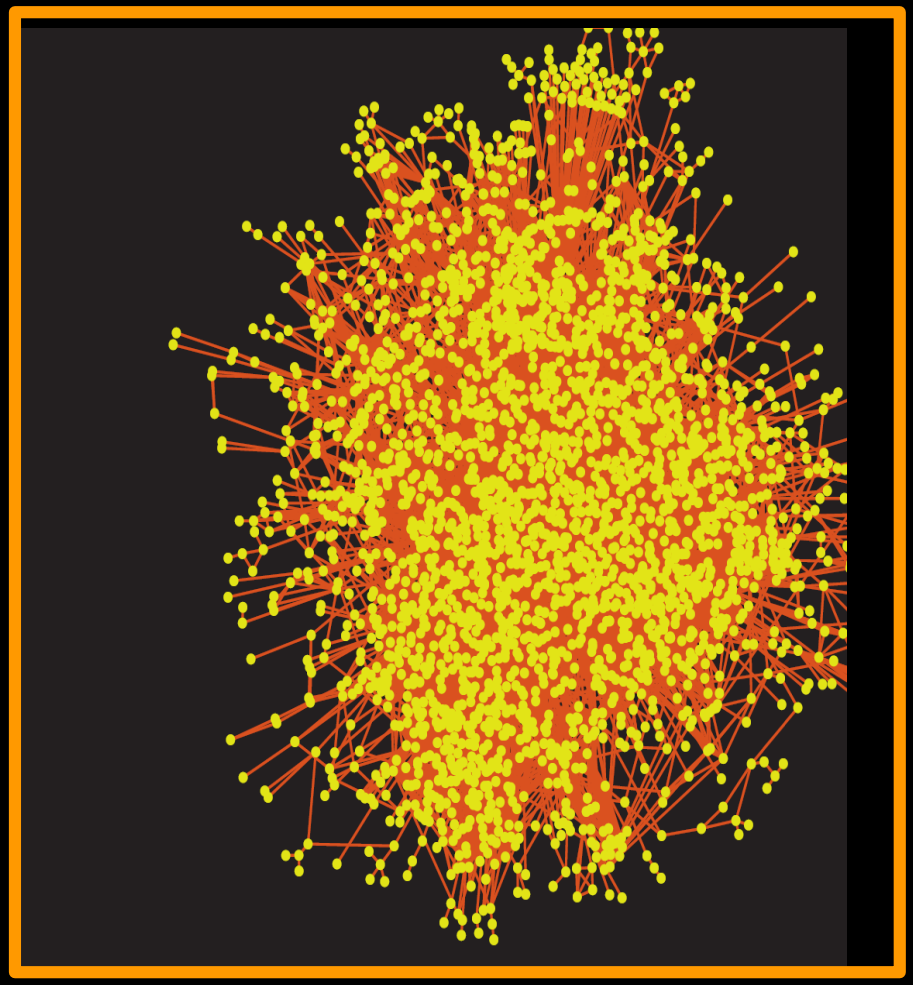
NIH / NLM / NCBI

Type of information we use

- Patient DNA information – (somatic) mutations and copy number variation of genes (done by comparing with normal tissues from the same patient)
- Gene expression: estimates “activity” of the gene.
- Other other markers – metabolites, epigenetic, ….
- Pathology description
- Survival time
- Age, sex

We aslo have a rough knowledge of interactions between genes





Key challenges in cancer data analysis

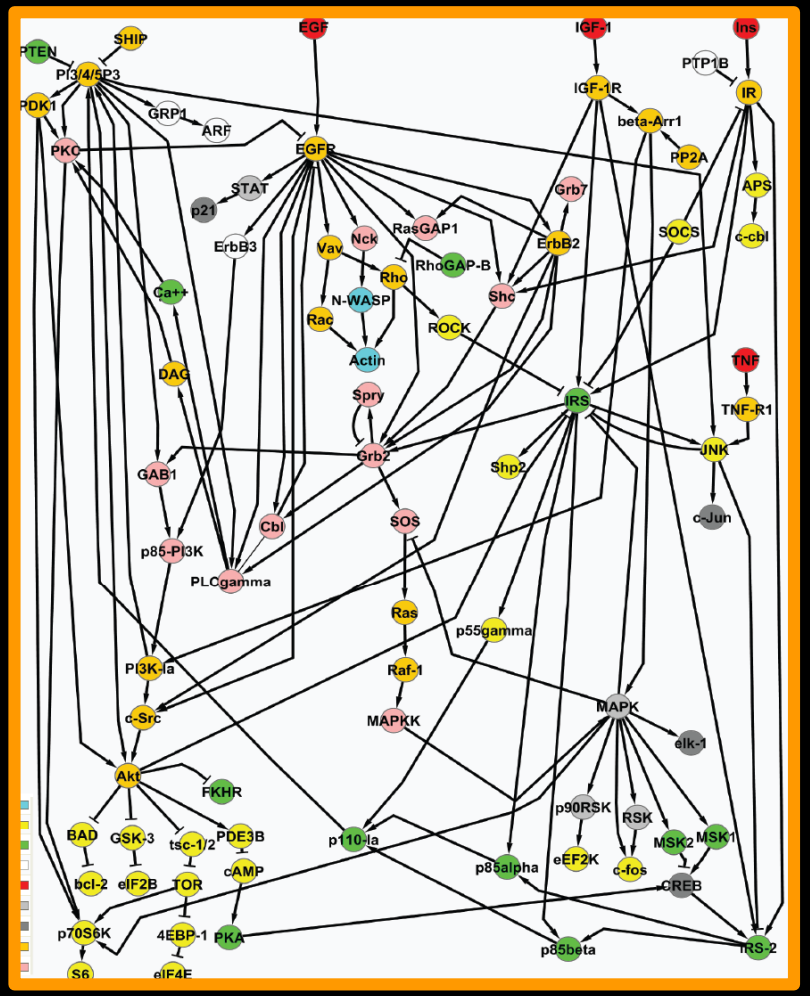
- Complexity: Multiple driver mutations are typically required for caner progression
- Heterogeneity: Phenotypically similar cancer cases might be caused by different sets of driver mutations
    - Driver mutations – mutations contributing to cancer progression
    - Passenger mutations – neutral mutations accumulating during cancer progression
- Some driver mutations are rare
- Epistasis – masking of the effect of one mutation by another mutation
- Cancer evolution

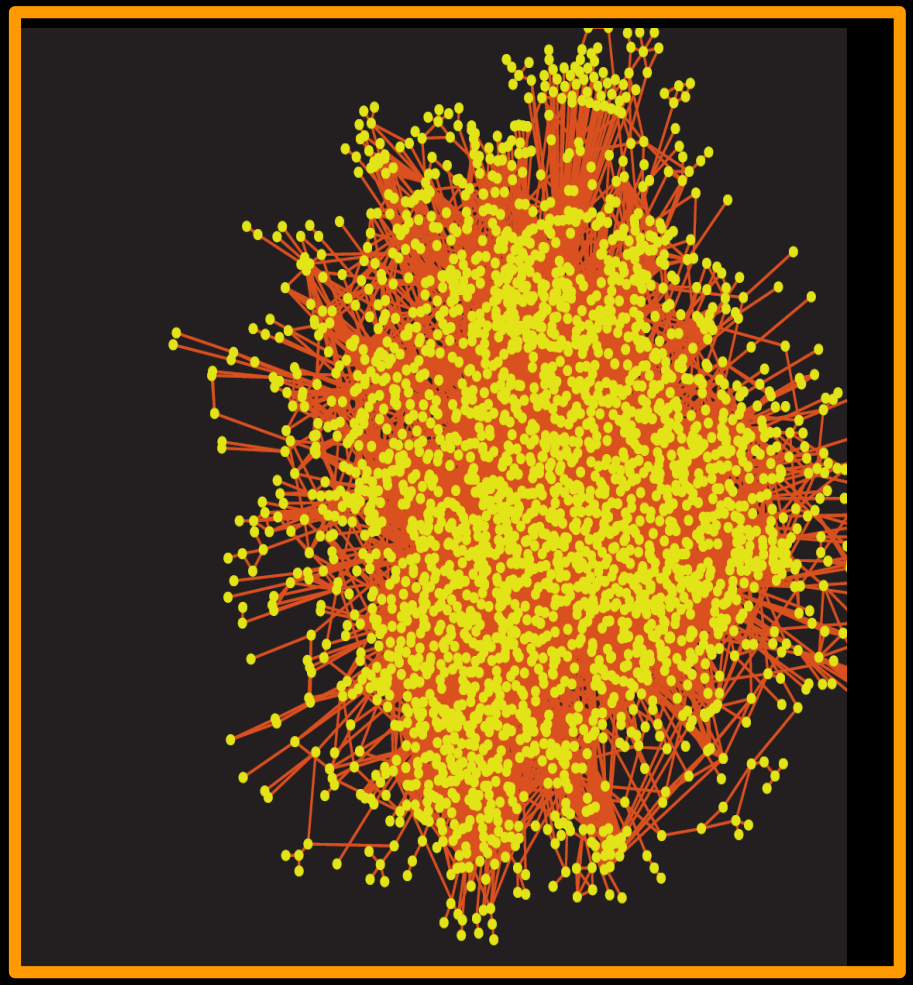
Network/Systems biology view

Motivation:

- Effects of genetic alteration propagate trough the interaction network affecting downstream genes
- Different driver mutations often dysregulate common pathways

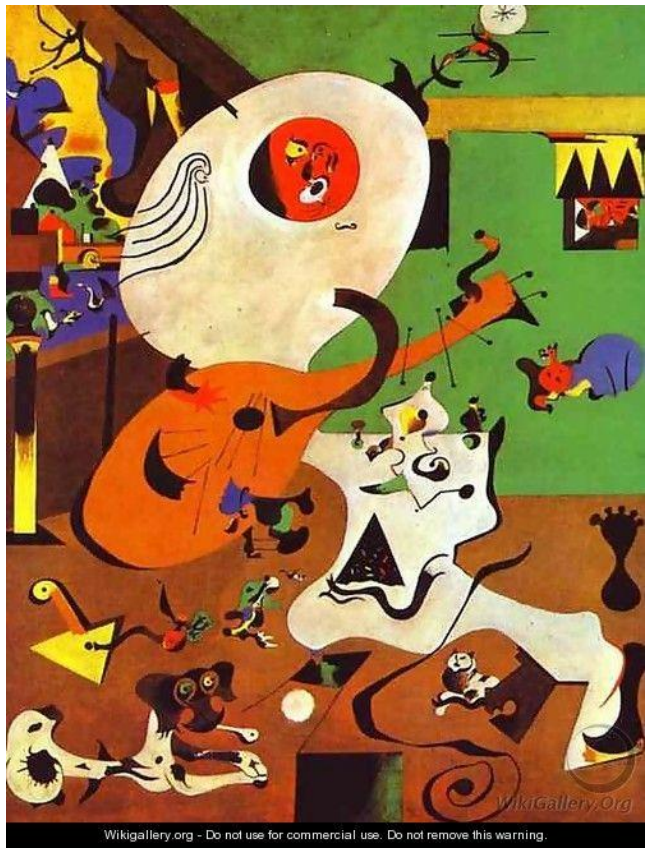
Which network to use?





High throughput network versus “the true” network

The Lute Player, Hendrick Maertensz Sorgh (1610-1670), Rijksmuseum, Amsterdam
(public domain)



Dutch Interior 1, Joan Miró (1893–1983)
Museum of Modern Art, New York
© 2012 Successió Miró / Artists Rights Society (ARS), New York / ADAGP, Paris
(used with ARS permission)

Kim, Przytycka, Frontiers 2013

First line of attack….

finding dysregulated pathways

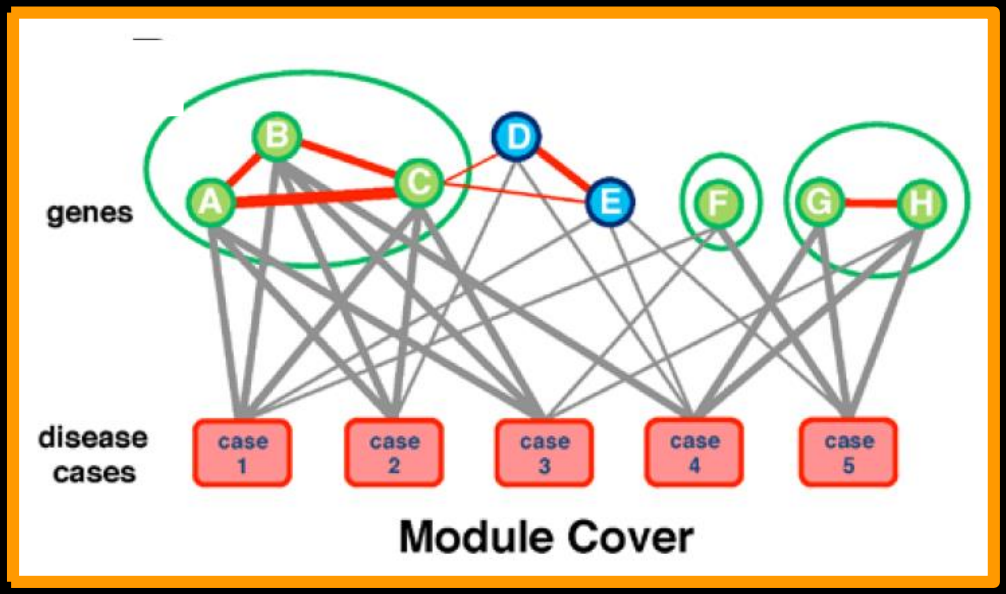
Module Cover Approach

Optimization problem:

Find smallest cost set of modules so that each disease case is covered at least k times

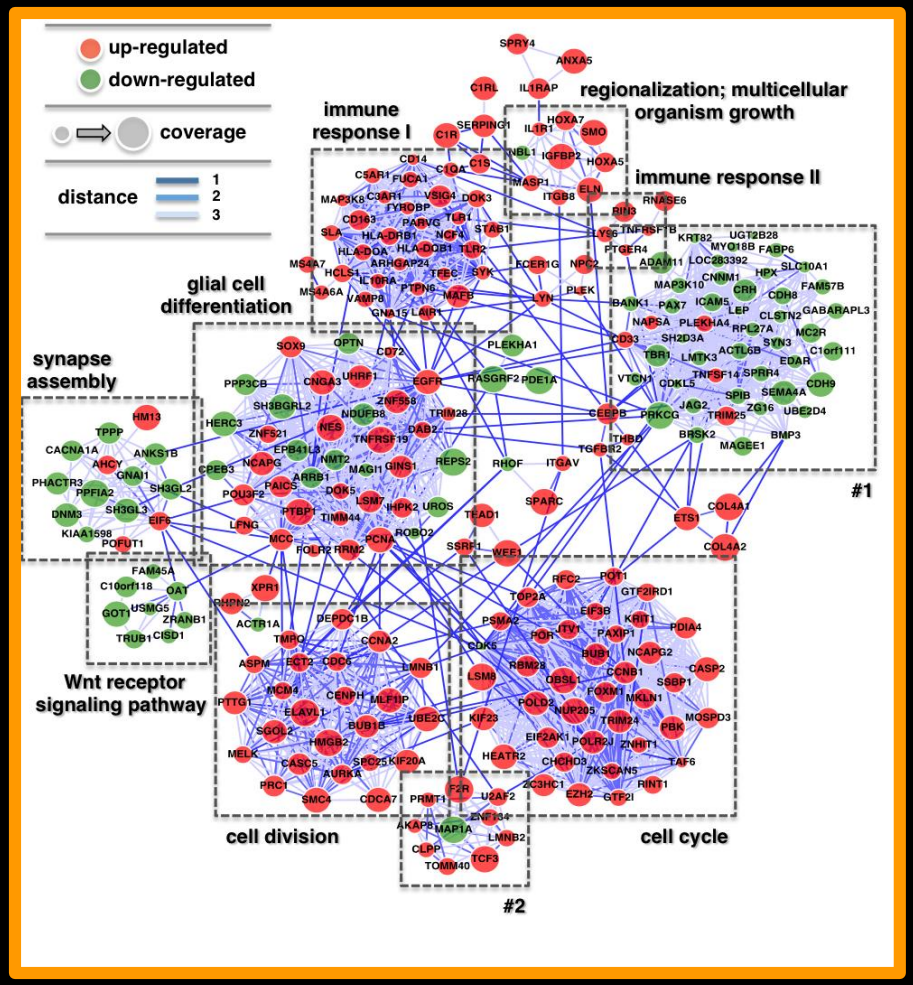
Cost is a function of:

- ↓ **distance in the network of genes in same module**
- ↓ **A similarity measure (application dependent)**
- ↑ **number of modules (parameterized penalty)**



Kim et al. PSB 2013

Module Cover: Glioblastoma Data

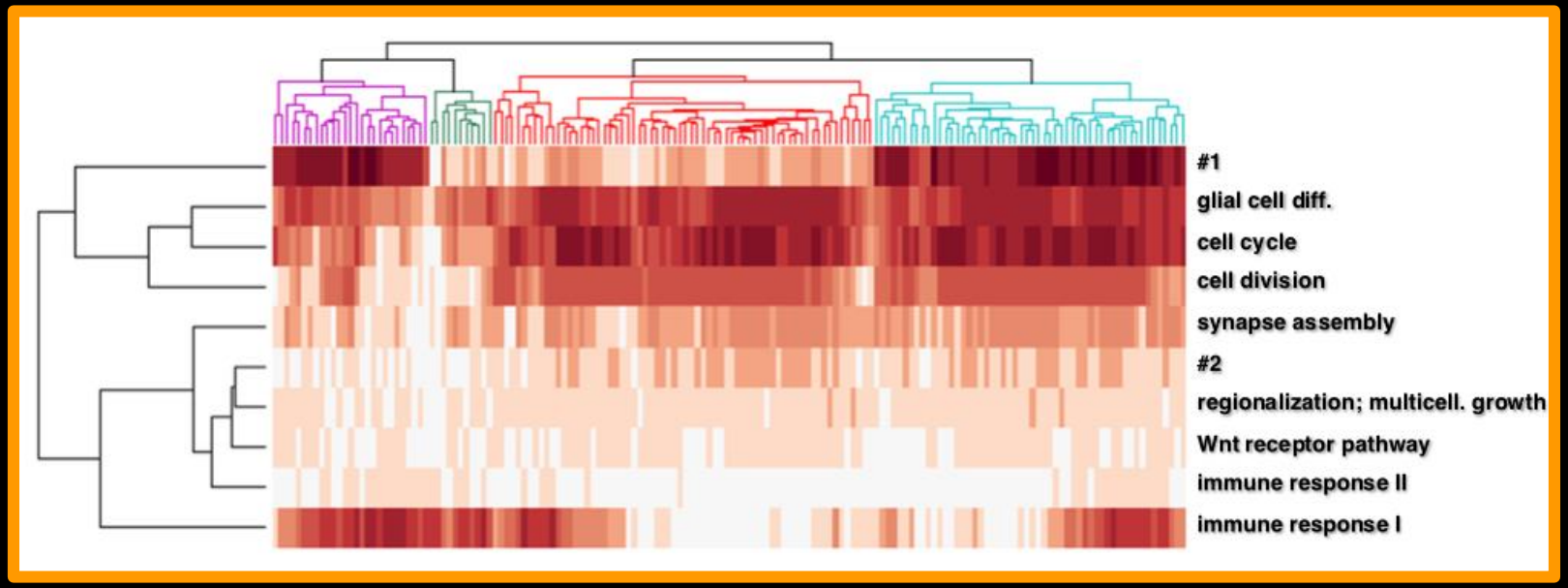


Signature modules from GBM Dataset (REMBRANDT)

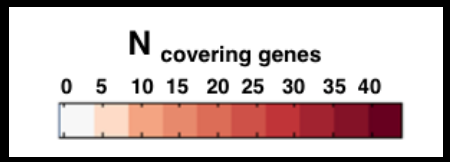
Kim et al. PSB 2013

Different patients groups have different signature modules

cases



modules



Kim et al. PSB 2013

From the recognition of heterogeneity to modeling

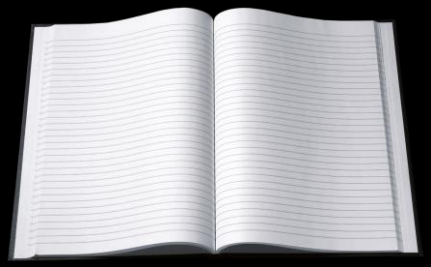
Second line of attack:

Using mixture models for capturing heterogeneity

Topic Model

Topic I

president 0.45
parliament 0.37
debate 0.21
….



Topic II

police 0.51
instigation 0.29
search 0.17
body….

Topics III

score 0.38
win 0.22
ball 0.14
….

Topic IV

Audience 0.27
Screen 0.15
Movie ….

Chang J, Blei DM: Hierarchical Relational Models for Document Networks. Ann Appl Stat 2010, 4(1):124-150.

Topic Model

Topic I

president 0.45
parliament 0.37
debate 0.21
….

Topic II

police 0.51
instigation 0.29
search 0.17
….

Topics III

score 0.38
win 0.22
upset 0.14
….

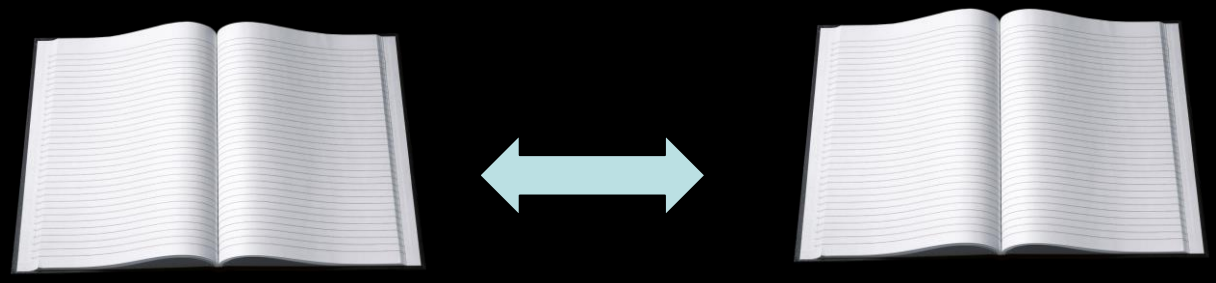
Topic IV

Audience 0.27
Screen 0.15
….

Chang J, Blei DM: Hierarchical Relational Models for Document Networks. Ann Appl Stat 2010, 4(1):124-150.

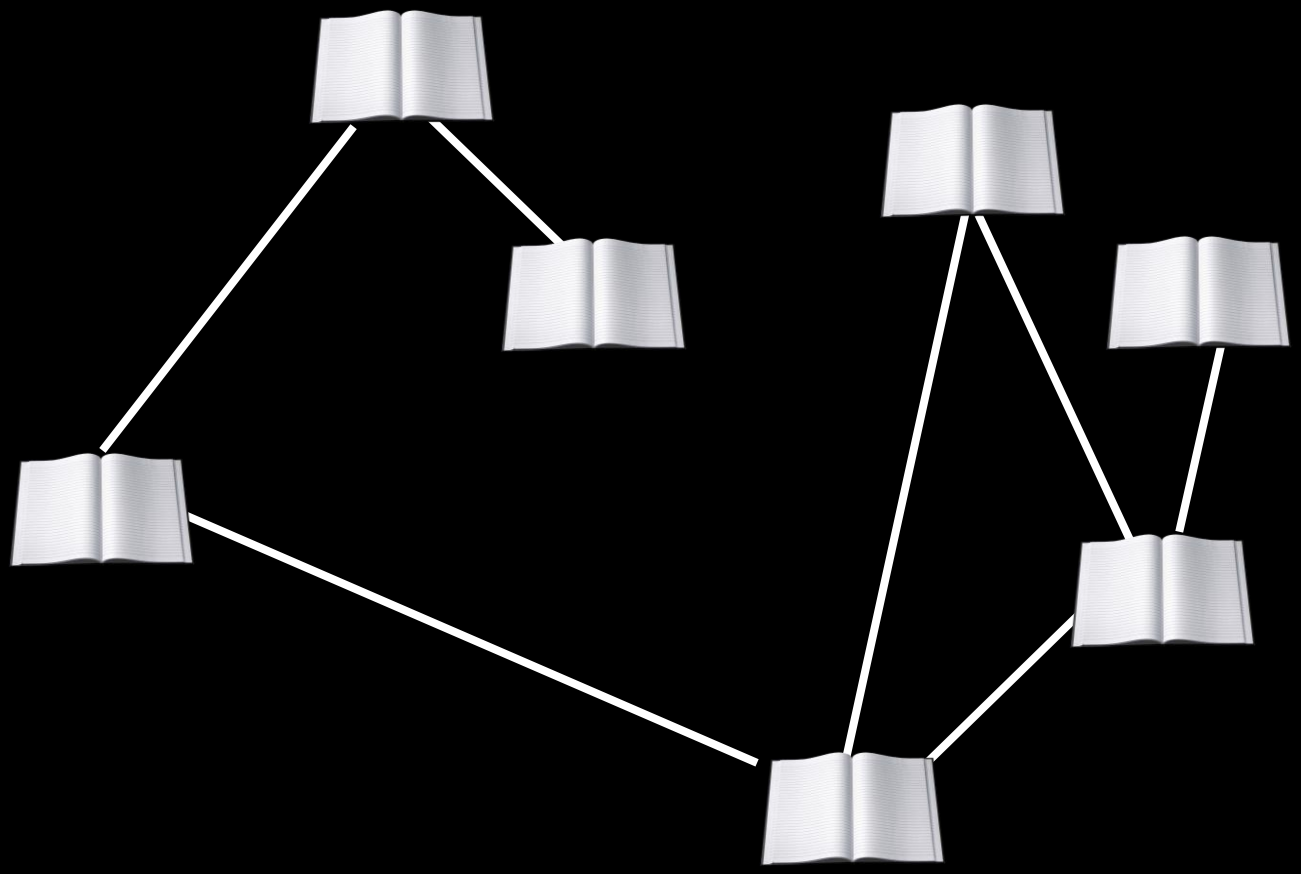
Additional information

Similar ?



Chang J, Blei DM: Hierarchical Relational Models for Document Networks. Ann Appl Stat 2010, 4(1):124-150.

Document similarity network



Chang J, Blei DM: Hierarchical Relational Models for Document Networks. Ann Appl Stat 2010, 4(1):124-150.

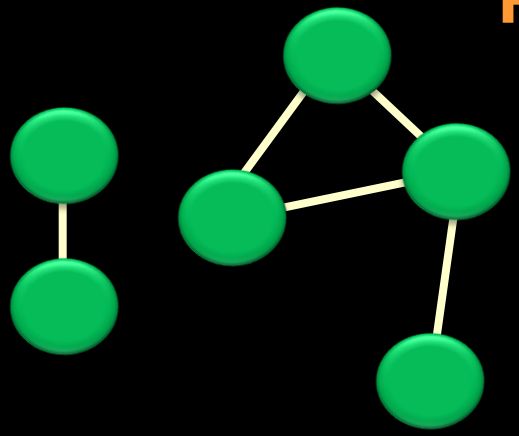
Topic Model for Cancer data

Explanatory features (words):

- mutations, CNV, micro RNA level;
- Epigenetic factors,
- Sex, age, environment ….

Phenotypic features

Survival time
Response to drugs,…..
Gene expression profile



Patient graph

Nodes – patients

Edges – phenotypic similarities

Key idea

neighbors in patient network should have similar explanatory features

Cho et al. NAR 2013/RECOMB 2012

Represent each patient as mixture of the subtypes

Subtype I

EGFR_A 0.45
NF1_M 0.37
PTEN_A 0.21
….

Subtype II

PDGFA_A 0.51
IDH1_M 0.29
M53_M 0.17
….

Subtype III

mirR218_H 0.38
ICDK2_D 0.22
SHC1_M 0.14
….

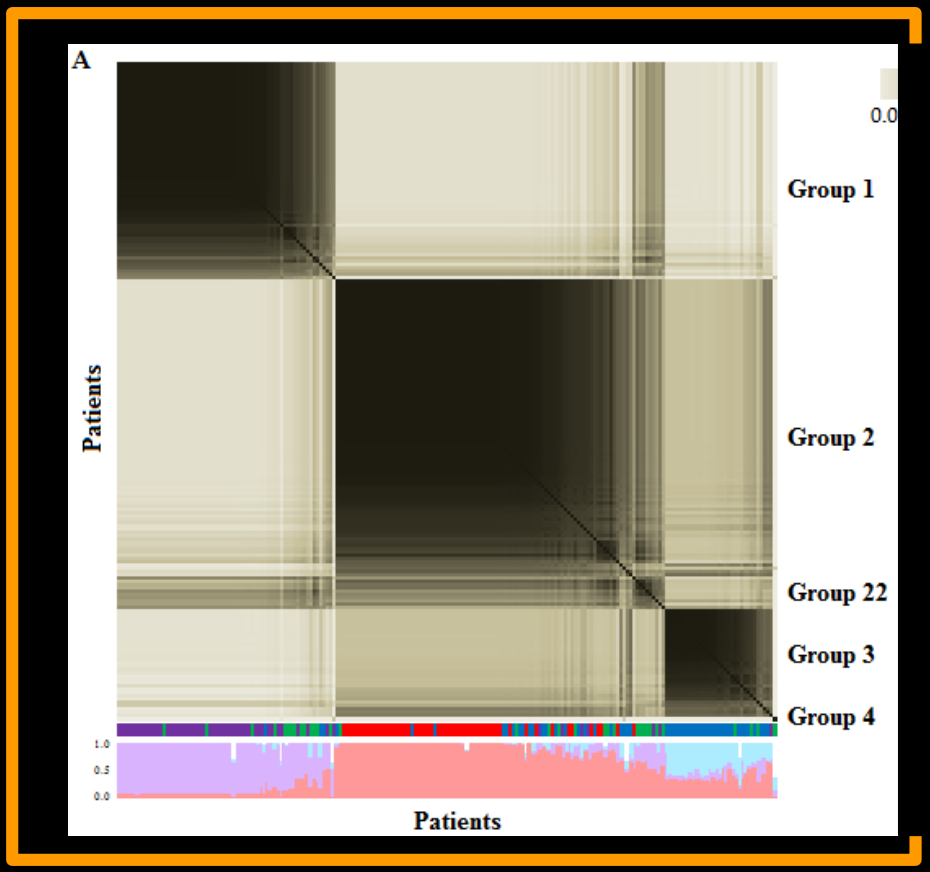
Features:
EGFR_A
NF1_M
CDKN2B_D
.
.
;

Subtype IV

CDK2B_D 0.37
EGFR_A 0.25
….

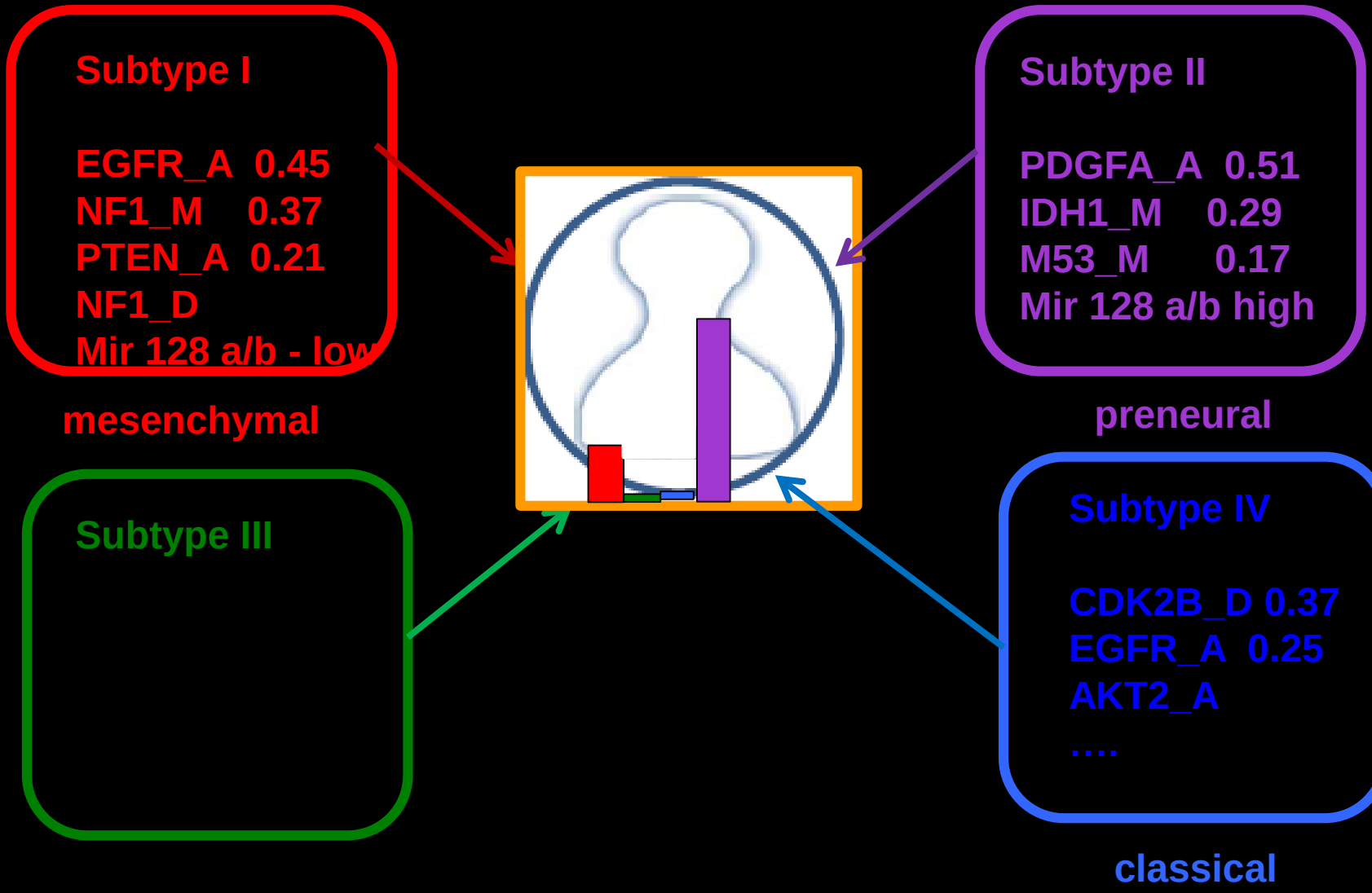
Cho et al. NAR 2013/RECOMB 2012

Patient-patient relationship based on1000 models



Cho et al. NAR 2013/RECOMB 2012

Connecting “causes (“words”) to subtypes



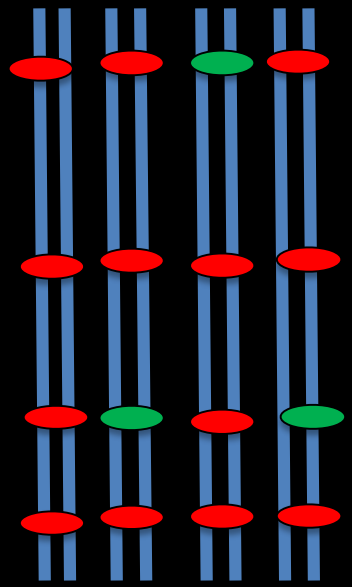
Cho et al. NAR 2013/RECOMB 2012

Third line of attack….

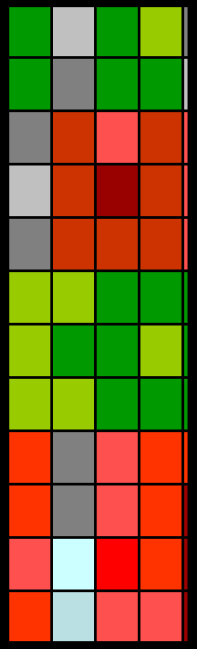
Information flow from
genotype to phenotypes

Information flow from genotypic changes to expression changes

Copy number aberrations or/and mutations

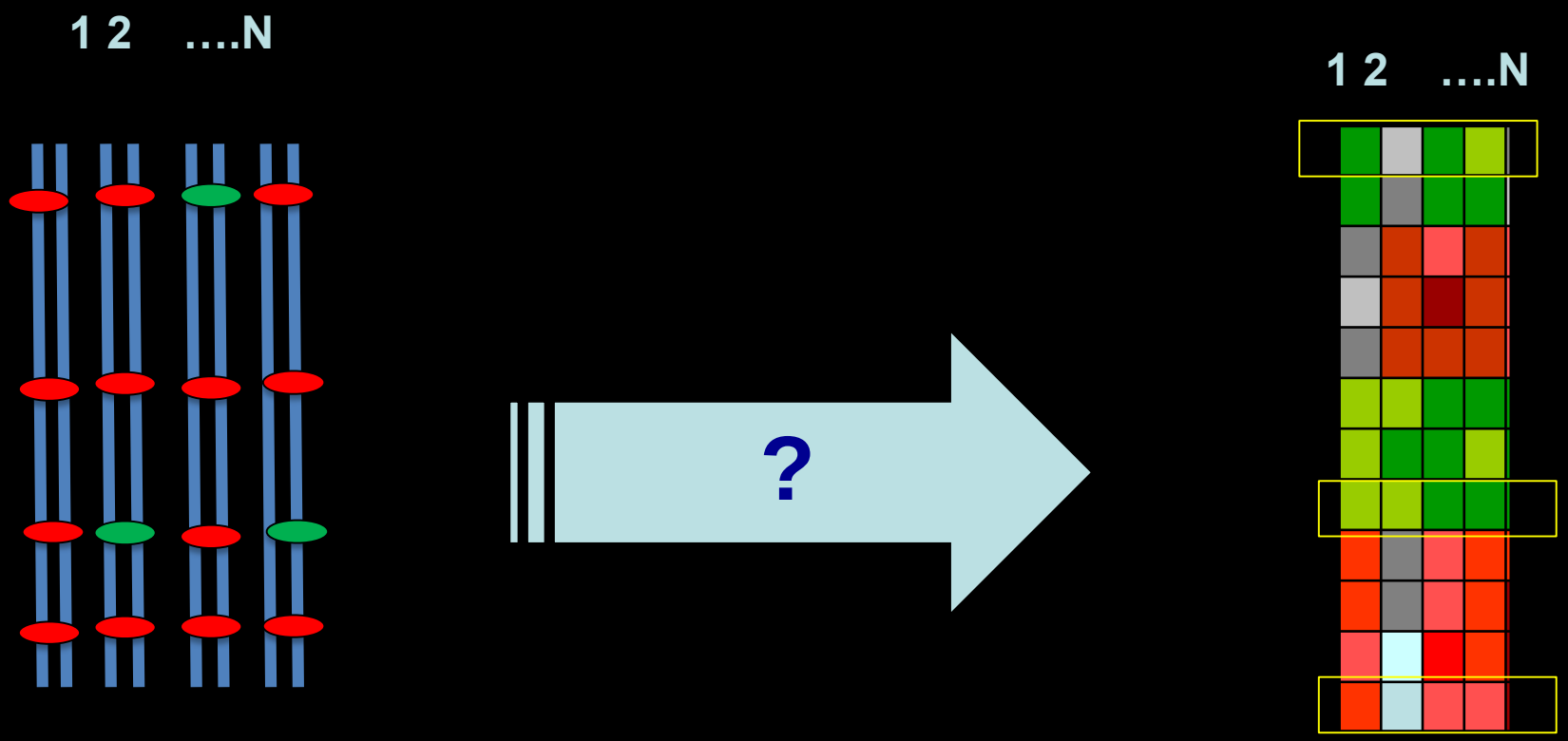


Gene expression

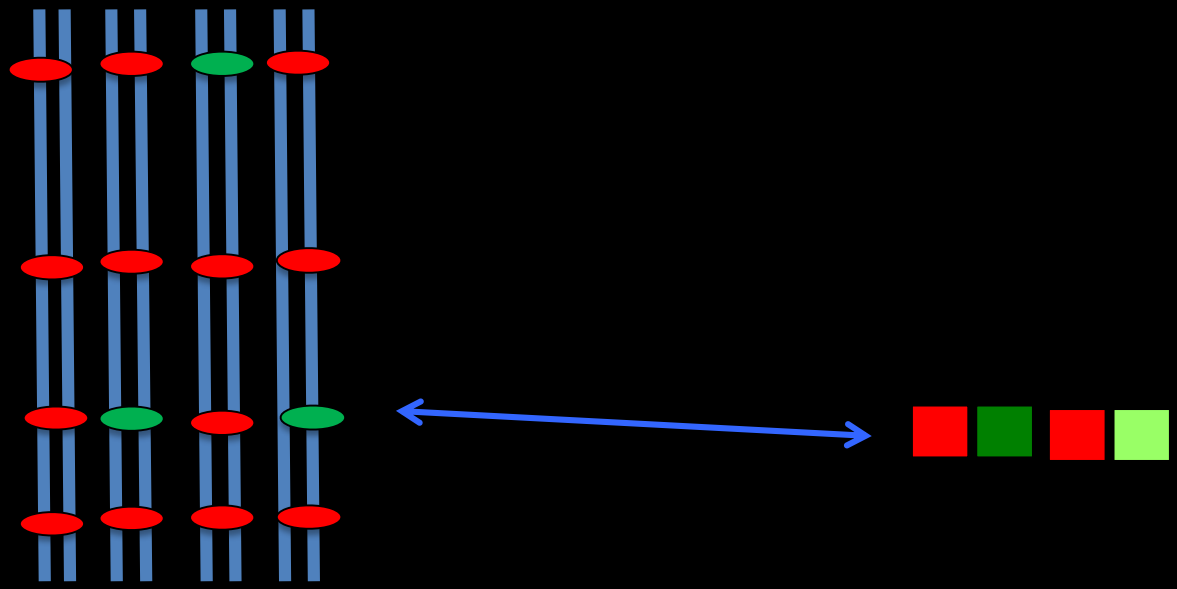


Kim et al. PolS CB 2011/RECOMB 2010

Explaining expression changes in the signature genes

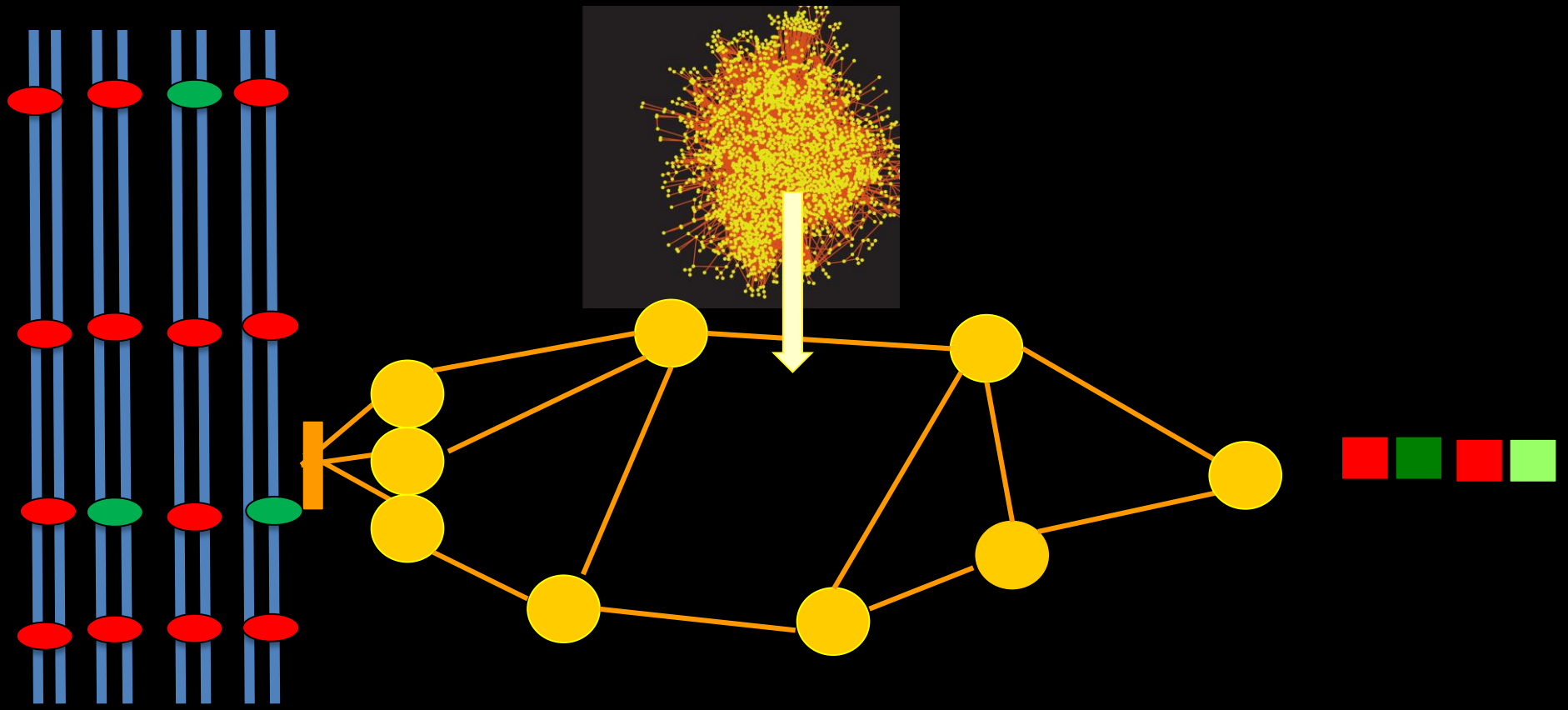


eQTL analysis links expression variability to genotypic variability



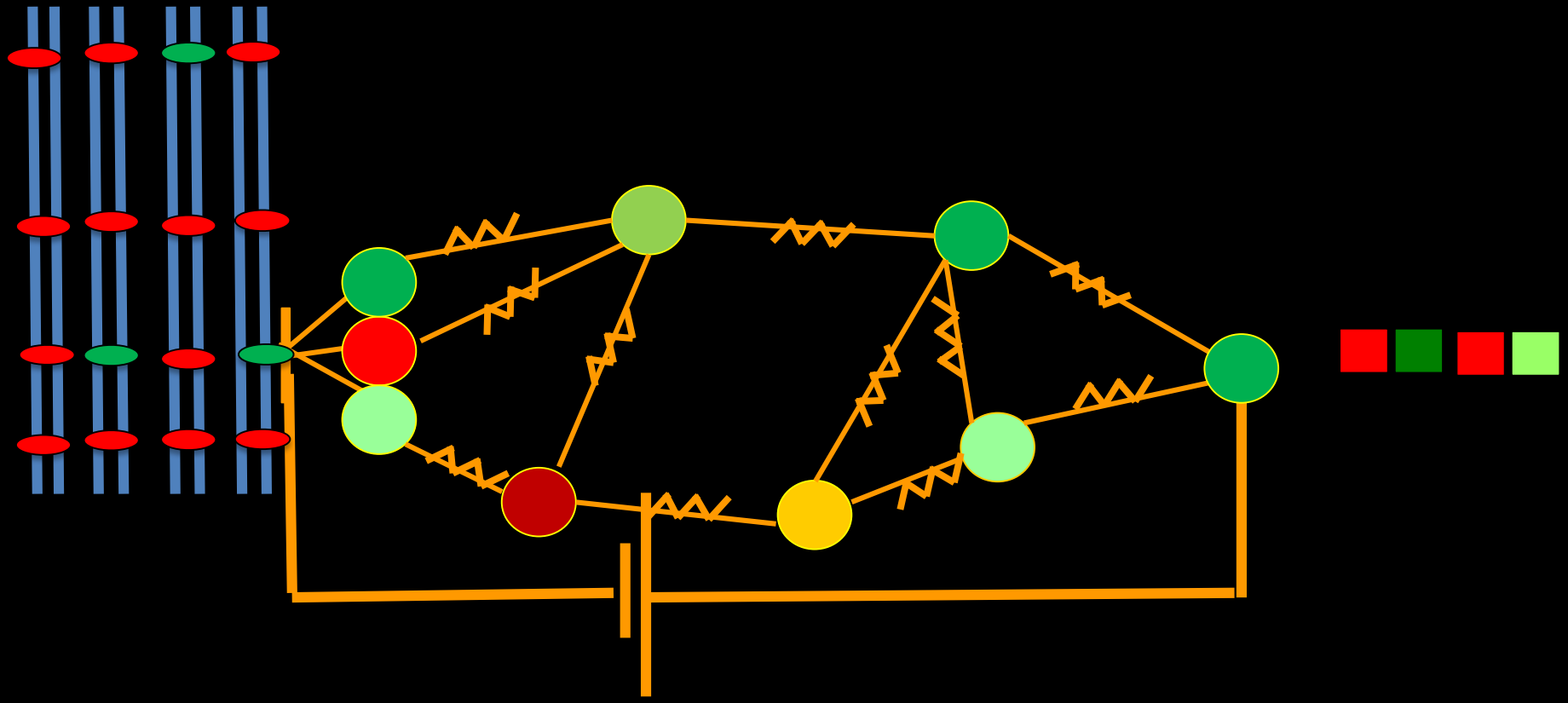
Tu *et al* Bioinfomatcis 2006
Suthram *et al* MSB 2008
Kim et al. PolS CB 2011/RECOMB 2010

Uncovering pathways of information flow between CNV and target gene



Tu *et al* Bioinfomatcis 2006
Suthram *et al* MSB 2008
Kim et al. PolS CB 2011/RECOMB 2010

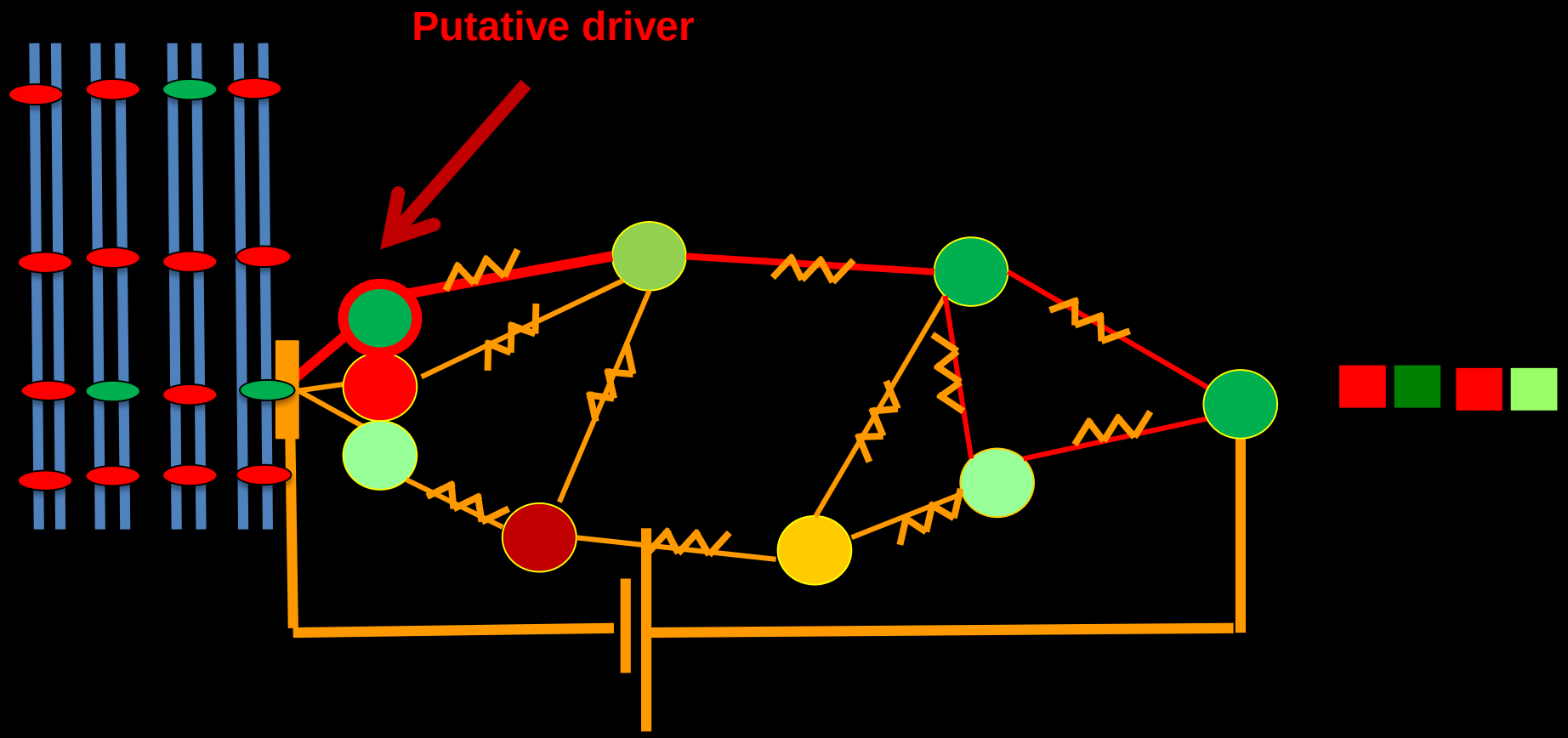
Adding resistances differentiate likelihoods of the edges



Resistance - set to favor most likely path -based on gene expression values

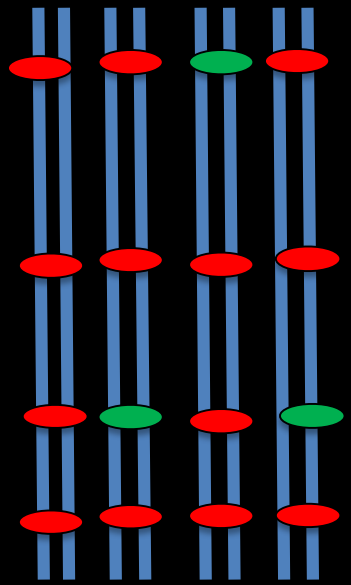
(reversely proportional to the average correlation of the expression of the adjacent genes with expression of the target gene)

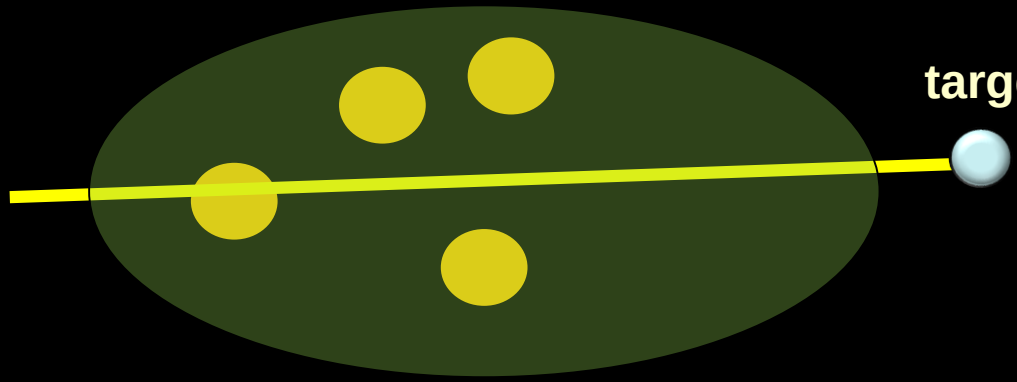
Finding subnetworks with significant current flow



Resistance - set to favor most likely path -based on gene expression values

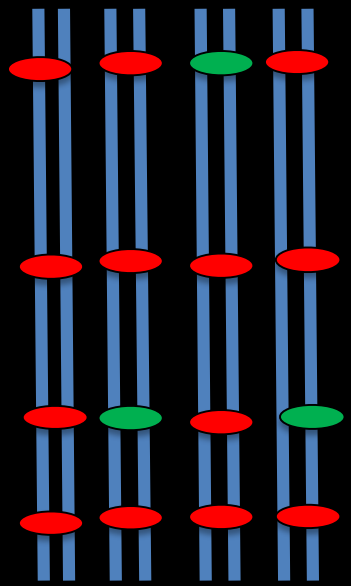
(reversely proportional to the average correlation of the expression of the adjacent genes with expression of the target gene)

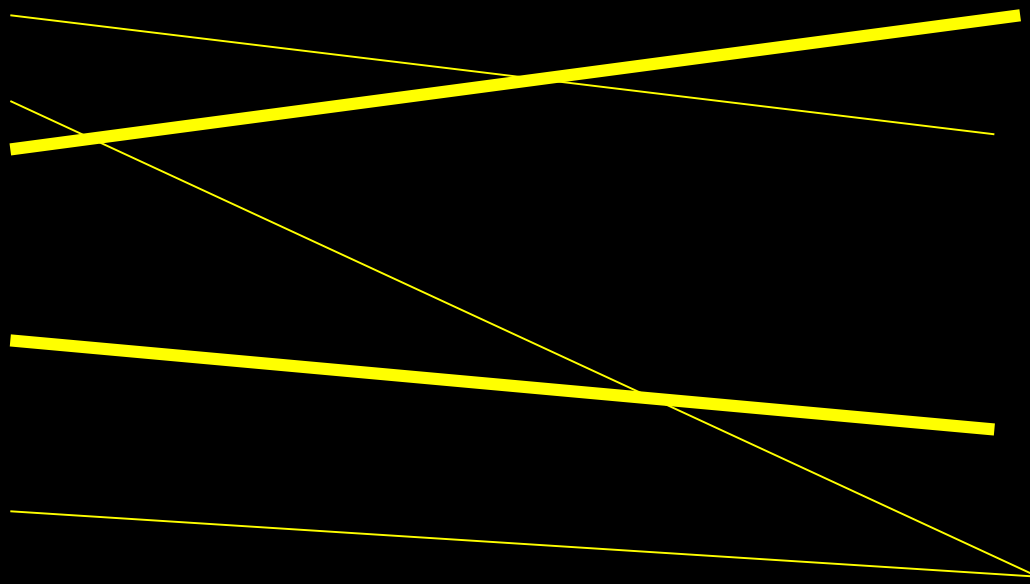


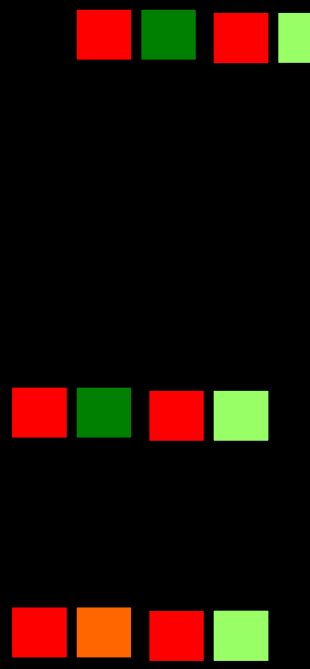


Kim et al. PolS CB 2011/RECOMB 2010

Repeat for other genes and significantly associated loci





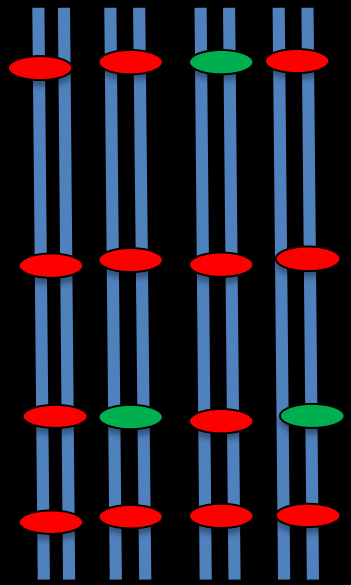


Cancer Cases
CNV data

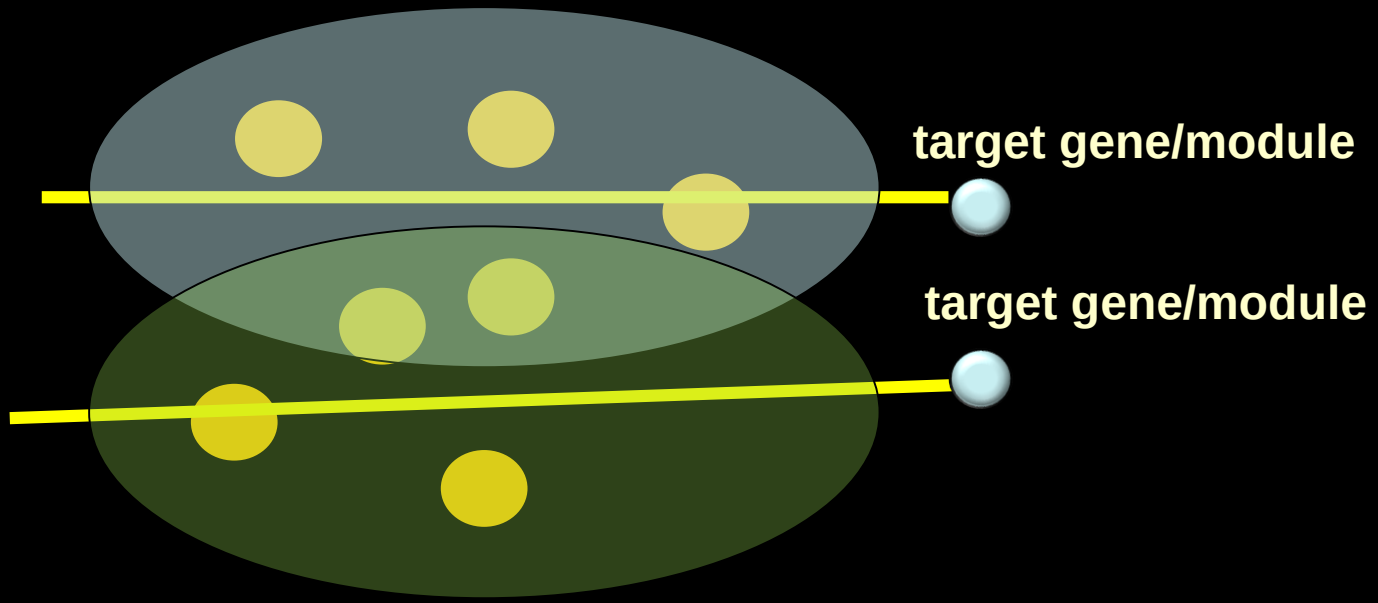
Cancer Cases
Gene expression data

Are there common functional pathways?

Cancer Cases
CNV data



Cancer Cases
Gene expression data

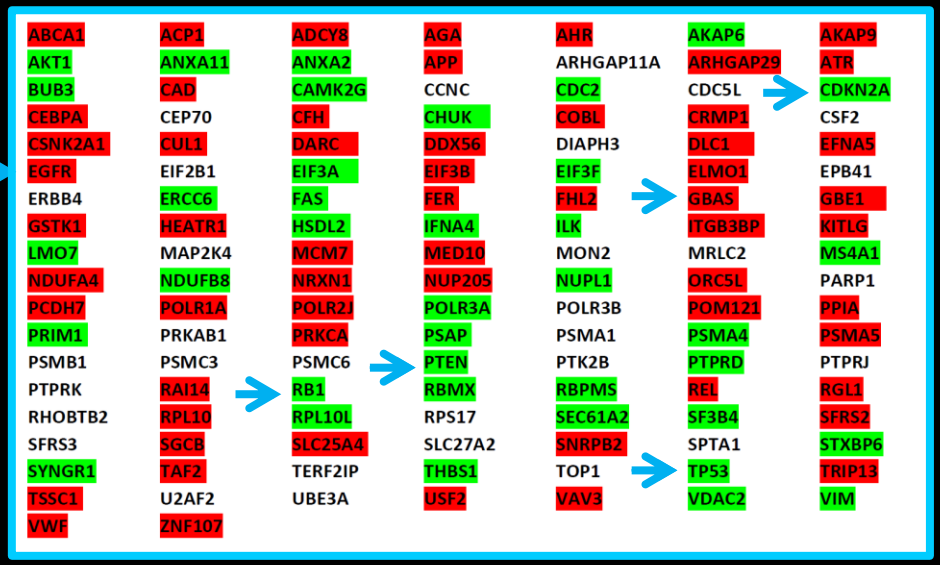


Kim et al. PolS CB 2011/RECOMB 2010

Gene Hubs

MYC(110)	E2F1(88)	E2F4(43)	CREBBP(34)	GRB2(27)	SP3(26)	ESR1(25)
TFAP2A(25)	NFKB1(23)	MYB(22)	JUN(22)	E2F2(22)	RELA(21)	AR(21)
SP1(20)	RPS27A(20)	MAPK3(19)	POU5F1(17)	HIF1A(16)	PPARA(15)	CDC42(15)
UBA52(13)	CDK7(13)	YBX1(13)	YWHAZ(12)	CEBPB(12)	POU2F1(12)	UBE2I(11)
SMAD3(11)	TAL1(11)					

Driving Copy number aberrations



Pathway Hubs

GO biological process	#
cell cycle arrest	10
epidermal growth factor receptor signaling pathway	9
negative regulation of cell growth	9
Ras protein signal transduction	9
regulation of sequestering of triglyceride	8
cell proliferation	7
nuclear mRNA splicing, via spliceosome	7
regulation of cholesterol storage	7
nucleotide-excision repair	7
RNA elongation from RNA polymerase II promoter	7
insulin receptor signaling pathway	6
transcription initiation from RNA polymerase II promoter	6
N-terminal peptidyl-lysine acetylation	5
phosphoinositide-mediated signaling	5
positive regulation of lipid storage	4
positive regulation of specific transcription from RNA polymerase II promoter	3
positive regulation of epithelial cell proliferation	3
base-excision repair	2
negative regulation of hydrolase activity	2
gland development	2
positive regulation of MAP kinase activity	2
regulation of nitric-oxide synthase activity	2
estrogen receptor signaling pathway	2
regulation of receptor biosynthetic process	2
response to organic substance	2
JAK-STAT cascade	2
regulation of transforming growth factor-beta2 production	2
G1/S transition of mitotic cell cycle	2
SMAD protein nuclear translocation	2

Summary

- Uncovering Cancer Heterogeneity trough data integration and network models
  - Module Cover method for finding dysregulated pathways
  - Topic model for cancer subtypes and their mixtures and determining features
  - Information flow from genotype to phenotype

AlgoCSB

Algorithmic Methods in Computational and Systems Biology





DongYeon Cho **Phung Dao** **Jan Hoinka** **YooAh Kim** **Damian Wojtowicz**

Collaborators:

Aptamers
Eli Gliboa (UM), Zuben Sauna (FDA); Marit Nilsen-Hamilton (U. Iowa)

DNA structural dynamics and its regulatory role
David Levens (NCI), Rafael Casellas (NCI)

Systems biology approach to gene dosage compensation
Brian Oliver (NIDDK)

Systems biology of cancer
Stefan Wuchty, Jozef Przytycki